

Second memo to D. Smith Mary



INTEROFFICE CORRESPONDENCE

DATE: September 14, 1993

TO: D. M. Smith, Remediation Programs Management, Bldg. 080, X8636

FROM: M. A. Siders, Geosciences, Bldg. 080, X6933
R. P. Boan, Geosciences, Bldg. 080, X8658 *AS*

SUBJECT: DRAFT ASSESSMENT OF THE IMPACT OF GILBERT'S STATISTICAL RECOMMENDATIONS -
MAS-007-93

The Gehan Test

We would like to repeat our apprehension in using the Gehan test for Operable Unit (OU) versus background comparisons. The Gehan test would be (by far) the most difficult test for subcontractors to use in their OU versus background comparisons. The test requires custom code in SAS, an expensive statistical software package that subcontractors generally do not have nor use. Most subcontractors have more popular commercial software, such as StatGraphics, and some use spreadsheets for their statistical work.

Perhaps more critical, however, is whether the Gehan test is appropriate for OU versus background comparisons. Helsel (1990) notes that, "In the most comprehensive review of these scores tests, most of them were found inappropriate for the case of unequal sample sizes." Gilbert himself cautioned us about the use of the untested and unproven Gehan test. The Gehan test was proposed as the way to deal with multiple detection limits. However, there are better ways to approach this problem such as providing the laboratories with better instructions regarding reporting of the data. This clear instruction will help to eliminate the "CRDL syndrome" - the main cause of data problems concerning the detection limit.

Comments on Background Study

You state that the background study did not address (1) suitability of background, (2) use of random sampling techniques, and (3) identification of spatial and temporal trends in background data. These statements are incorrect. The 1992 Background Plan and the 1992 Background Report both detail the methodology used in (1) selecting the most suitable background sites, (2) describes the placement of sampling sites as random as practically feasible, and (3) testing for seasonality or other variations over time (1989-93).

Impacts on Schedule and Budget

Gilbert recommended using the 99/99 UTLs as a "hot-spot" test; this should reduce (not increase) the number of analytes flagged as potential COCs, when compared with the 95/95 UTLs now used for essentially the same purpose. It is critical to realize that Gilbert did not suggest using all five

DOCUMENT CLASSIFICATION
REVIEW WAIVER PER
CLASSIFICATION OFFICE

Attachment E

D. M. Smith
September 14, 1993
MAS-007-93
Page 2

statistical tests for any one set of data comparisons; rather, he suggested "tandem testing" (using two tests) plus a "hot-spot" test. The schedule and cost impact (omitting OUs 1 and 2) should be slight. What will save time and money is giving the subcontractors clear instruction on how to treat non-detects, how to perform data cleanup, and how to proceed with the OU versus background comparisons.

Data Cleanup Issues, Non-Detect Replacement

During our work on the 1993 Background Report, we tested the efficacy of various ways of treating non-detects and performed "tandem testing" (t-test and Wilcoxon Rank Sum test) for the nonparametric ANOVA used in the Background Report itself. We discovered that, for as much as 50 to 80 percent non-detects, simple substitution (NDs replaced with 1/2 result) gave nearly the same results as using Helsel's method of replacing non-detects. The simplicity and ease of using simple substitution more than makes up for any decrease in power. Helsel (1990) cautions against using heavily censored data sets "...especially for legal or management purposes...." Helsel (1990) also notes that when severe censoring occurs, all tests "...have little power to detect differences in central values."

In short, it is probably unwise to base decisions on very heavily censored data sets (greater than 80 percent non-detects). Simple substitution is nearly as "correct" as more complex methods (Helsel's, Cohen's, etc.) of replacing non-detects. Non-detects should never be dropped from the data set. Both Sanford et al. (1993) and Helsel (1990) stress this.

General Comments

It is important to note that there is no "Gilbert method" per se. Gilbert merely reviewed different types of statistical tests that *may be appropriate* to use in OU versus background comparisons. The reason he offered different tests was because he realized that some tests work better than others for a given data set.

Also, your remark under "II Technical Aspects" stating that, "The Gilbert method (see previous paragraph) is generally quite conservative in that its application will likely minimize the chance of missing site contaminants at the expense of increasing the likelihood of falsely declaring analytes as contaminants when in fact they are not" is not correct. Gilbert's proposals are conservative, but there is *no indication* that the implementation of Gilbert's proposal will increase the likelihood of falsely declaring analytes as contaminants.

jlm

cc:

M. E. Levin *oib*

of knowledge below the reporting limit. Results do not depend on a distributional assumption (25).

When severe censoring (near 50% or more) occurs, all of the above tests have little power to detect differences in central values. The investigator will find it difficult to state conclusions about the relative magnitudes of central values, and other characteristics must be compared. For instance, contingency tables (class 3) can test for a difference in the proportion of data above the reporting limit in each group (20). This test can be used when the data are reported only as detected or not detected. It also may be used when response data can be categorized into three or more groups, such as below detection, detected but below some health standard, and exceeding standards. The test determines whether the proportion of data falling into each response category differs as a function of different explanatory groups, such as different sites or land use categories.

Hypothesis testing with multiple reporting limits. More than one reporting limit often is present in environmental data. When this occurs, hypothesis tests such as comparisons between data groups are greatly complicated. The fabrication of data followed by computation of *t* tests or similar parametric procedures is at least as arbitrary with multiple reporting limits as with one reporting limit, and should be avoided. Also, data below all reporting limits should never be deleted before testing.

Tobit regression (class 2) can be used with multiple reporting limits. Data should have a normal distribution around all group means and equal group variances to use the test. These assumptions are difficult to verify with censored data, especially for small data sets.

One robust method that always can be used is to censor all data at the highest reporting limit, and then perform the appropriate nonparametric test. Thus the data set

<1 <1 <1 5 7 8 <10 <10 <10 12 16 25
would become

<10 <10 <10 <10 <10 <10 <10 <10
<10 12 16 25

and a rank-sum test would be performed to compare this with another data set. Clearly, this causes a loss of information which may be severe enough to obscure actual differences between groups (a loss of power). For some situations, however, this is the best that can be done.

Alternatively, nonparametric score tests common in the medical "survival analysis" literature sometimes can be applied to the case of multiple reporting limits (26). These tests modify uncensored rank test statistics to compare groups of data. The modifications allow

for the presence of multiple reporting limits. In the most comprehensive review of these score tests (27), most of them were found inappropriate for the case of unequal sample sizes. Another crucial assumption of score tests is that the censoring mechanism must be independent of the effect under investigation (see box). Unfortunately, this often is not the case with environmental data. The Peto-Prentice test with an asymptotic variance estimate was found to be the least sensitive to unequal sample sizes and to differing censoring mechanisms (27).

In summary, robust hypothesis tests have several advantages over their distributional counterparts when they are applied to censored data. These advantages include freedom from adherence to a normal distribution; greater power for the skewed distributions common to environmental data; comparisons between central values such as the median, rather than the mean; and the incorporation of data below the reporting limit without fabrication of values or bias. Information contained in less-than values is used accurately and does not misrepresent the state of that information.

When adherence to a normal distribu-

tion can be documented, tobit regression (class 2) offers the ability to incorporate multiple reporting limits regardless of a change in censoring mechanism. Score tests (class 3) require consistency in the censoring mechanism with respect to the effect being tested.

Methods for regression

With censored data, the use of ordinary least squares (OLS) for regression is prohibited. Coefficients for slopes and intercept cannot be computed without values for the censored observations, and substituting fabricated values may produce coefficients strongly dependent on the values substituted. Four alternative methods capable of incorporating censored observations are described below. The first and last approaches, Kendall's robust fit (28) and contingency tables (20), are nonparametric (class 3) methods requiring no distributional assumptions. Robust correlation coefficients also are mentioned (20). Tobit and logistic regression (24, 29), the second and third methods, fit lines to data using maximum likelihood (class 2). Both methods assume normality of the residuals, though with logistic regression, the assumption is after a logit

The appropriateness of score tests

When a score test is not appropriate

Score tests are inappropriate when the censoring mechanism differs for the two groups. That is, the probability of obtaining a value below a given reporting limit differs for the two groups when the null hypothesis that the groups are identical is true.

1. Suppose a trend over time is being investigated. The first five years of data are produced by a method that has a reporting limit of 10 µg/L; the second five years of data are compiled by an improved method with 1 µg/L as its reporting limit. A score test of the first half of the data versus the second would not be valid because the censoring mechanism itself varies as a direct function of time.

2. Two groups of data are compared as in a rank-sum test, but most of the data from group A were measured with a chemical method having 1 as its reporting limit, and most of group B were measured with a method having 10 as its reporting limit. A score test would not yield valid results because the censoring mechanism varies as a function of what is being investigated (the two groups).

When a score test is appropriate

A score test yields valid results when the change in censoring mechanism is not related to the effect being measured. Stated another way, the probability of obtaining data below each reporting limit is the same for all groups, assuming that the null hypothesis of no trend or no difference is true. Here a score test provides much greater power than does artificially censoring all data below the highest reporting limit before using the rank-sum test.

1. Comparisons have been made between two groups of data collected at roughly the same time and analyzed by the same methods, even though those methods and reporting limits have changed over time. Score tests are valid in this case.

2. Differing reporting limits result from analyses performed at different laboratories, but each sample had been assigned at random to the different laboratories. Censoring thus is not a function of what is being tested, but is a random effect, and score tests would be valid.

Helsel, 1990
Attachment F